A Unified Perspective on Adversarial and Out-of-Distribution Detection in the Open World

Yeli Feng, Daniel Jun Xian Ng, Arvind Easwaran

Nanyang Technological University
50 Nanyang Ave, 639798, Singapore
yfeng002@e.ntu.edu.sg, danielngjj, arvinde@ntu.edu.sg

Abstract

Deep neural networks (DNN) have achieved near-human classification capability when testing samples are drawn from their training data distribution. However, numerous research also revealed that the performance of DNN can degrade severely when testing samples are maliciously manipulated or out of distribution (OOD). In response, research in both adversarial defense and OOD detection have become very prevalent independently. This paper investigates the interplay between these two approaches and attempts to unify them to increase the robustness of classification systems in the open world, where inputs could be in-distribution (ID), adversarial, OOD, or adversarial-OOD. We find that existing defensive training methods, adversarial or data augmentation based, trade classification and OOD detection performances for robustness to adversaries. We propose an algebraic transformation based data augmentation technique that reduces DNN's sensitivity to adversarial attacks. Furthermore, we formulate a multi-level semantics backed detection metric to enhance OOD detection capability by utilizing multi-task training. In combination, our defensive training method, SVrandom+, mitigates the trade-off between performance and robustness. In experiments our method achieved a true positive rate (TPR) of 89.2% for OOD detection and an error of 9.3% for classification in the open world setting. Furthermore, its TPR is higher by 16.7%, and classification error is lower by 2.5 times than existing gradient based adversarial training.

Acknowledgment

This research was funded in part by MoE, Singapore, Tier-2 grant number MOE2019-T2-2-040.

Introduction

Deep Neural Network (DNN) based intelligent systems have been rapidly deployed in the real world both as standalone software and components in cyber-physical systems. Many of them, such as autonomous driving and biometric authentication, are safety or security critical. Among DNN perception systems, object classification is a very important, if not the central, task. Inevitably, it will face unknown or novel objects in the real world, namely OOD samples, hence it is very likely to give wrong outputs in such situations. Detecting and rejecting OOD samples is fundamental to increasing DNN's robustness in the open world. Over the years, many detection methods have been proposed by the OOD community. Some of them utilize the object classifier outputs as OOD features, while others look into generative learning techniques (e.g., Hendrycks and Gimpel 2016, Zenati et al. 2018, Hsu et al. 2020).

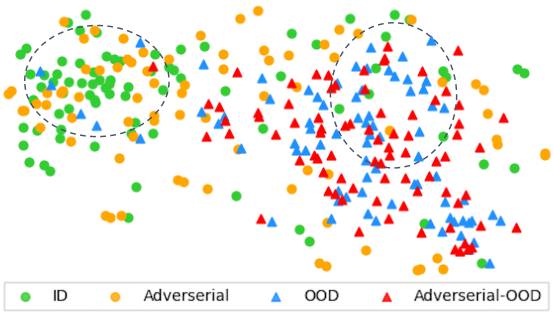

Figure 1: A view of classification in the open world from its logits-space that is reduced to two dimensions by the t-SNE method (Van der Maaten and Hinton 2008). ID and OOD samples are from benchmark datasets CIFAR10 and SVHN respectively. Adversarial and adversarial-OOD samples are generated from them. However, all samples are predicted as dog class.

Orthogonally, Szegedy et al. (2013) demonstrated that a DNN can be easily fooled to make prediction errors. An example is when perturbations, hardly perceivable to the human eye, are injected into benign images; the DNN can make such errors even with very high confidence. This discovery raised many concerns and has since led to active research for methods attacking (e.g., Moosavi-Dezfooli et al. 2017, Madry et al. 2018) or defending DNNs (e.g., Madry et al. 2018, Shafahi et al. 2020). So far, adversarial training is one of the most effective defense techniques among input transformation, randomization, model ensemble, and certified defense approaches (Dong et al. 2020).

For DNN classifiers in the open world, inputs could be ID, adversarial, OOD, or adversarial-OOD, as exemplified in Figure 1. And not to mention unintentionally corrupted images. Hence, it is seemingly advantageous to integrate the adversarial training's defense capabilities with the rejection abilities from OOD detectors into a single unified framework. In doing so, an aggregated robustness benefit might be

achieved for classification systems in the open world. Some recent studies have begun to explore this interplay between OOD detection and adversarial defense. Methods designed for OOD detection showed promising lab performance in detecting adversaries (Lee et al. 2018). However, existing adversarial training introduces a performance trade-off between classification robustness and accuracy (Zhang et al. 2019), as well as between robustness and OOD detection capability (Song et al. 2020).

In this paper, we extend the investigation along the aforementioned vein. We found that the trade-off is not just limited to gradient-based adversarial training. Other existing defensive training methods, such as some input transformation and data augmentation based methods, also trade robustness to adversaries for classification and OOD detection performances.

We propose an algebraic input transformation based method to reduce the DNN's sensitivity to malicious perturbations without the performance trade-off in both classification and OOD detection tasks. Specifically, we treat the linear components (singular subspaces) produced by singular value decomposition (SVD) like learned representations in the latent space of generative learning, and reconstruct robustness-oriented training samples from such singular subspaces. Utilizing multi-task learning technique to obtain an extra set of logits as OOD features, we subsequently propose a multi-level semantics backed OOD detection metric that enhances OOD detection performance.

We selected three ID domains and conducted evaluations over five types of adversarial attacks and five to six OOD domains for each ID domain. Compared to three existing defensive training methods, experiments showed our method, SVrandom+, overcomes the aforementioned trade-off problem. Purposing OOD detection as a fail-safe controller for classifiers in the open world setting, at 10% FPR, our method achieves a TPR that is 15.0% to 30.1% higher on CIFAR10, and 14.8% to 35.7% higher on CIFAR100. Details see in Table 3.

Related Work

Adversarial Defense. Training DNN with adversarial samples perturbed by well-designed adversarial attack methods is an intuitive approach. One of the first methods along this direction used the fast gradient sign attack (FGSM) to add adversarial samples to the training process (Goodfellow, Shlens, and Szegedy 2014), where its objective function is a weighted summation of loss over ID samples and their adversaries.

Inspired by the success of the empirical risk minimization (ERM) principle in finding classifiers with small population risk, Madry et al. (2018) incorporate adversaries into ERM's definition of population risk. The authors formulated a training objective that minimizes the population loss over model parameters, meanwhile maximizing the allowed perturbation power. The authors further proposed a more powerful attack method, projected gradient descent (PGD), and showed the combination significantly improved robustness to a wide range of attacks. However, the computation cost of the PGD attack method is very high. To overcome this, Wong, Rice, and Kolter (2019) took the min-max formulation but replaced the PGD attack with an iterative FGSM with a random initialization trick. The authors showed that their method is as effective as the above mentioned PGD-based training but significantly reduces training time. Shafahi et al. (2020) also took the min-max recipe, but applied the universal perturbation proposed by (Moosavi-Dezfooli et al. 2017) for adversaries.

As one of the most effective defense approaches, these methods tend to bear a non-trivial trade-off between classification accuracy and robustness to adversaries. Moreover, the iterative gradient searching process for adversaries also contributes to prolonged training times.

Another adversarial defense approach relies on input transformation to detect or eliminate perturbations in the malicious samples. In the past, many detection-based defense methods have been proposed, such as bit-reduction, jpeg compression, and feature squeezing. Dong et al. (2020) benchmarked a wide range of adversarial defense methods and found that the input transformation-based defenses slightly improve robustness over undefended classifiers. Applying the halftoning operation to convert images into a density-based representation during training and testing, Huang, Liao, and Huang (2021) reported that their method can achieve effective defense with some trade-off in performance.

Recent data augmentation techniques have been found helpful in enhancing the classifiers' performance and its generalization ability. Inspired by the Vicinal Risk Minimization (VRM) principle, Zhang et al. (2018) proposed an augmentation technique that produces more virtual samples around the distribution boundary of training data by mixing up images and their target classes. The authors showed that among other benefits such as reduced memorization of corrupt labels, their method reduces the DNN's sensitivity to adversarial samples. Gong et al. (2021) proposed to augment data with lightweight Gaussian perturbation, which in theory, introduces an extra gradient-norm regularization to a standard training objective. Compared to using iterative gradient searching methods to generate adversaries from inputs, data augmentation incurs a much less computation cost.

Multi-class OOD Detection. Multi-class classifiers use the maximum value of the posterior class probabilities' output ($p(y|x)$) from a softmax function for prediction, as given in Equation 1, where T is a temperature parameter that equals to 1 in standard softmax classifiers, $f_i(x)$ is the logit of the predicted class i and C is the total number of classes.

$$S_i(x;T) = \max_i \frac{e^{f_i(x)/T}}{\sum_{j=1}^{C} e^{f_j(x)/T}} \tag{1}$$

Hendrycks and Gimpel (2016) found that the probabilities of correctly predicted ID samples are often sufficient for detecting OOD samples and proposed an OOD detection metric that works with any DNN softmax classifier, denoted the *baseline* henceforth. It has been known that the softmax probabilities give a biased high confidence when the logits $f_j(x)$ are small across all classes. Therefore, by introducing a temperature scaling strategy and a small amount

of adversarial perturbation to test samples, ODIN enhanced detection performance significantly (Liang, Li, and Srikant 2018). A drawback of ODIN is that a specific T value has to be found for each OOD dataset to achieve optimal performance. Recently, Hsu et al. (2020) proposed a generalized ODIN that learns a calibration scale from inputs instead of tuning the hyperparameters. Lee et al. (2018) proposed a Mahalanobis distance based OOD metric that ensembles the DNN hidden layers' activations and softmax probabilities into OOD features per predicted class. The authors reported that their method effectively detects adversaries as well.

The increasing deployment of DNN into mission-critical systems in the open world have driven OOD detection into an active research field. More related literature can be found in a comprehensive survey (Bulusu et al. 2020).

Proposal

In this section, we propose two methods that aim to reduce prediction errors from classification systems deployed in the open world, where inputs could be ID, adversarial, OOD and adversarial-OOD. First, we describe a new data augmentation technique that utilizes outputs from SVD operation. It reduces the DNN's sensitivity to adversarial perturbations. We then introduce an enhancement to the *baseline* OOD detection metric through exploiting the parent-child relationship in semantic concepts.

Singular Subspaces based Reconstruction

In linear algebra, SVD operation factorizes a rectangular matrix into a canonical form in eigenvalues matrix $\mathbf{\Sigma}$ and orthonormal eigenbasis matrices $\mathbf{U}$ and $\mathbf{V}$, as given in Equation 2. On the diagonal of $\mathbf{\Sigma}$ are singular values. The columns of $\mathbf{U}$ and $\mathbf{V^T}$ are left and right singular vectors, respectively. When $\mathbf{X}$ represents image pixels, each pair of singular vectors encode the structure of an image layer and the corresponding singular value specifies its luminance. Each $u_i \sigma_i v_i^t$ forms a singular subspace, and the total number of subspaces equals the width of an image.

$$\mathbf{X} = \mathbf{U\Sigma V^T} = \tag{2}$$

$$\begin{bmatrix} | & | & \\ u_1 & u_2 & \cdots \\ | & | & \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \cdots \\ 0 & \sigma_2 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} | & | & \\ v_1^t & u_2^t & \cdots \\ | & | & \end{bmatrix}$$

The σs on the diagonal are in descending order. The most dominant geometry in an image can be reconstructed from the first singular subspace $u_1 \sigma_1 v_1^t$. See an example in Figure 2(c). SVD has been widely used in image processing tasks such as compression, denoising, and watermarking (e.g., Andrews and Patterson 1976, Guo et al. 2015, Chang, Tsai, and Lin 2005). However, we found it to be underappreciated in the deep learning field for image problems. For example, SVD is absent in a recent survey on data augmentation techniques (Shorten and Khoshgoftaar 2019).

SVD decomposes an image into linearly independent components. Each singular subspace encodes some geometric information in the image. Naturally, we can treat these singular subspaces as features, in the sense of learned features in the latent space of generative learning, and reconstruct variants in numerous permutations of luminance (singular values) and geometric information (singular vectors). As the goal here is to reduce sensitivity to adversaries, we devise a reconstruction strategy that randomly drops singular subspaces with trivial information (reduction), while on the other hand add trivial information from other images (swap). Reduction functions the same as denoising, while swap interpolates geometric information between a pair of images. Figure 2(d) shows an example of swap, where the pattern of the water surface is perturbed by a visually non-dominant structure in the pizza image. During training, randomness can be injected on the fly to control the degree of smoothing and interpolation, as described in Algorithm 1.

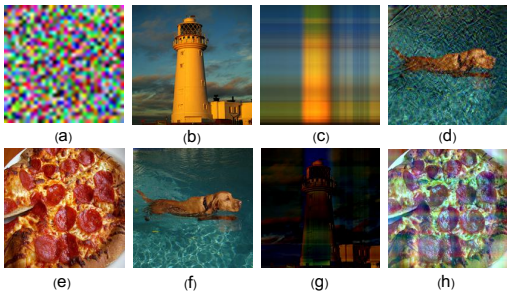

Figure 2: Image Reconstruction Examples
Image (a) is Gaussian noise. (b),(e) and (f) are original images from ImageNet. (c) is a reconstruction from the top-1 singular subspace of (b) and (g) is a reconstruction from the remaining subspaces. (d) is a reconstruction after swapping the bottom-k singular subspaces between (f) and (e). (h) is a reconstruction after the same swap operation but between (e) and (a).

Algorithm 1: Proposed data augmentation algorithm

Input: A mini batch of training images X
Parameter: A list specifies augmentation options and *limit* values
Output: Reconstructed image batch $\tilde{X}$

Randomly pick one option, reduction or swap
if swap **then**
  for $x_a, x_b \in X$ **do**
    Draw a random k uniformly from [0, *limit*]
    Swap bottom k singular subspaces from outputs of SVD operation over x_a and x_b
    Generate $\tilde{x_a}$ and $\tilde{x_b}$ from swapped matrices
    Add $\tilde{x_a}$ and $\tilde{x_b}$ to $\tilde{X}$
else
  for $x \in X$ do
    Draw a random k uniformly from [*limit*, *image-width*]
    Generate $\tilde{x}$ from top 1 to k singular subspaces of x
    Add $\tilde{x}$ to $\tilde{X}$
end if
return $\tilde{X}$

Multi-level Semantics for OOD Detection

One of our goals here is to investigate the interplay between adversarial defense and OOD detection, with a focus on defensive training. We need to use one simple yet

effective OOD detection metric for evaluation. The *baseline* (Hendrycks and Gimpel 2016) is ideal due to its simplicity and effectiveness. By simplicity, we mean that the *baseline* uses softmax outputs directly as OOD scores without any hyperparameter tuning and post-learning step. Detection methods using softmax outputs as OOD features would benefit principally the same way. Thus, we can focus on studying the impact of different defensive training methods.

Experimentally, we observed that when a classifier's accuracy is not sufficiently high, the OOD detection performance of a classifier trained with the proposed data augmentation method is lower than its counterpart that is trained without any defense. Using CIFAR100 as an example, its training set includes 100 classes but 500 images only for each class. The best classification accuracy that can be achieved now is still below 90%. However, these 100 classes can be consolidated into fewer classes by their child-parent semantic relationships; for example, a dog is a kind of animal. If an image cannot be correctly classified by its parent concept, a correct prediction by its child concept can be seen as a source of uncertainty. Otherwise, it violates the interlinked conceptual-semantic relationship that we use to understand the world.

Based on the above idea, we propose to utilize multi-task learning technique (Zhang and Yang 2021) to obtain an extra logit to increase total discriminate power on OOD features. The intuition here is that it will be harder for adversarial perturbations to successfully alter two predictions over one image in such a way that they satisfy a correct child-parent relationship as described in the Appendix. Specifically, a classifier now has two tasks formulated by Equation 3a, where y^p is a parent concept of label y, and $\alpha \in [0, 1]$ weighs the loss between the two tasks.

$$\min_{\theta} \quad \alpha\mathcal{L}(\theta, x, y) + (1-\alpha)\mathcal{L}(\theta, x, y^p) \tag{3a}$$

$$\text{oodscore} = \alpha f_i(x) + (1-\alpha) f_{i^p}(x) \tag{3b}$$

In the *baseline*, the detection threshold is determined by the softmax probabilities of test ID samples that are correctly classified. With this multi-level semantics design, we expand the *baseline* so that the detection threshold is determined by test ID samples correctly classified by both the target and the corresponding parent labels. Specifically, our OOD score is a weighted summation over logits of predicted child and parent classes as defined in Equation 3b. Another change we introduce to the *baseline* is replacing the softmax probabilities with its inputs, i.e., the logits, considering that the softmax function will incur a bias towards high confidence when logit values are small across all the classes.

We found that simply using an α value around the accuracy of the corresponding single task classifier improves overall robustness. Details see in comparison and ablation studies in the next section.

Evaluation

We proposed a unified approach to increase classifiers' robustness to adversarial, OOD and adversarial-OOD samples. Framed for practical settings, we evaluate the effectiveness of our proposal and compare it with related defensive training methods. Firstly, we explain the experiment setup. Secondly, we recommend a conservative benchmark that better aligns with the open world setting. Experiment results and performance are discussed subsequently. Finally, an ablation study is conducted to identify the contribution of each element in our proposal.

Experiment Setup

The novelty of our proposal lies in the training method, so we pick three related studies that take different approaches towards classification robustness for comparison. They are gradient-based adversarial training Fast-FGSM (Wong, Rice, and Kolter 2019), input transformation based adversarial training Halftone (Huang, Liao, and Huang 2021) and data augmentation based Mixup (Zhang et al. 2018).

For experiments, we used the datasets and ID/OOD definition as given in Table 1a. CIFAR, SVHN and LSUN are benchmark datasets from Torchvision. GTSDB (Houben et al. 2013) was selected as a training set so that our experiments covers ID domains with diversity in classification accuracy and image complexity. The GTSDB and LSUN images were resized to 32-by-32 pixels of the same size as all other images. Gaussian samples are random noise with variance equal to 1. Uniform samples are quasi-solid in color with pixel values ranging in $x \pm 1.01\%$, where $x \in (0, 255)$. Gaussian noise and quasi-solid color simulate two ends of extreme OOD distributions where pixel perturbation or luminance reduction renders semantics in natural images not perceivable to the human eye.

Table 1: Experiment Setup

ID	OOD
CIFAR10	SVHN, GTSDB, LSUN, Gaussian, Uniform
CIFAR100	SVHN, GTSDB, LSUN, Gaussian, Uniform
GTSDB	CIFAR100, CIFAR10, SVHN, LSUN, Gaussian, Uniform

(a) ID and OOD Sets

Method	CIFAR10	CIFAR100	GTSDB
Fast-FGSM	0.877 / 0.838	0.720 / -	0.865 / -
Halftone	0.805 / -	0.492 / -	0.935 / -
Mixup	0.934 / 0.958	0.739 / 0.789	0.985 / -
SVrandom+	0.953 / -	0.780 / -	0.980 / -

(b) Classification Accuracy (our-result / originally-reported)

For each ID training set, five classifiers were trained with methods listed in Table 1b. SVrandom+ means the classifier proposed in this study trained with the multi-level semantics. Fast-FGSM and Mixup were adapted from authors' implementation available in the public domain, where the Preact-Resnet18 network was used. In our implementation, ResNet101 was used for CIFAR100 and ResNet18 for CIFAR10 and GTSDB classifiers. Accuracy over corresponding ID test set from our training results are presented in the

table, along with the performance originally-reported by the respective studies when available. A random horizontal flip was applied to all methods in our implementation.

We trained all classifiers on PyTorch with the SGD optimizer and the cosine annealing method to schedule the learning rate, as in (Liu 2017). The base learning rate is 0.1, the weight decay is $5e^{-4}$ and training epochs are 200. CIFAR100 classifiers were further fine-tuned for 200 epochs. Given the difference in the network and training hyperparameters, our training results align with those reported in the original studies, except for the performance of *mixip* is lower on CIFAR100. We tested with the default mixing values (α) 1 and 0.1 from the range $[0.1, 0.4]$ recommended by the study to avoid under-fitting, but didn't obtain a result close to the originally reported. The mixing values used for the accuracy reported in Table 1b under our-result are 0.1 for CIFAR100 and 1 for the others. However, our Fast-FGSM model achieved a higher accuracy when compared to the originally reported number. Epochs achieving the best classification accuracy were used for analysis in all cases.

Blind-test for OOD Detection

The *baseline* popularized by Hendrycks and Gimpel (2016) adopts an ID-OOD paired setting (i.e., test samples in each evaluation are from either ID or a particular OOD domain). Detection performance measured in such paired tests could be too optimistic for the detection power actually achievable in the open world, as pointed out by several works on OOD detection benchmarks (e.g., Shafaei, Schmidt, and Little 2018, Ahmed and Courville 2020). In the open world, test samples could be unknown or from multiple known OOD domains. As outlined by the dotted line in Figure 3, a detection threshold that achieves close to 100% accuracy over the CIFAR10-Uniform pair will drop to about 80% over the CIFAR10-LSUN pair.

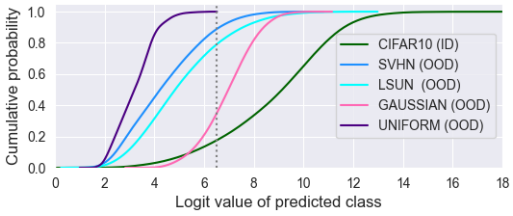

Figure 3: Distribution of OOD scores (logit values of predicted classes) over ID and various OOD test sets.

Therefore, we recommend a benchmark that relies only on the ID domain used for training a classifier. By fixing the false positive rate (FPR) (i.e., the degree of tolerance that ID test samples can be falsely flagged as OOD), we measure OOD detectors' true positive rate (TPR). Thus, the threshold setting mechanism is blind to any specific OOD domain. In other words, the popular paired-test gauges the performance upper bound of an OOD detection method. Whereas, our blind-test reports a relatively conservative outcome, if not a lower bound, that is achievable in the open world.

Comparison Studies

The analysis of the classifiers' robustness in the open world is organized into three groups of tests: 1) the adversarial defense capability, 2) the OOD detection ability, and 3) the overall performance using a combo test set comprising ID, adversaries, OOD and adversarial-OOD samples.

Adversarial Defense We use four white-box and one black-box attacks to evaluate the adversarial defense capabilities of the classifiers. Adversarial samples for white-box attacks were generated from the ID test sets using Torchattacks (Kim 2020). As listed in Table 2, these attacks are l_∞ FGSM and PGD, l_2 CW and Deepfool (Carlini and Wagner 2017, Moosavi-Dezfooli, Fawzi, and Frossard 2016). Default values in Torchattacks were used for adversarial parameters with the following exceptions: perturbation budget is 2/255 for FGSM, 2/255 with 2 iterations for PGD, and confidence is 1 for CW. Noise is a kind of black-box attack that is launched from random points in the image space. The perturbation budget used here is a uniform distribution of $\pm$10/255. For FGSM, PGD, and noise attacks, entire ID test sets were used to generate adversaries. CW and Deepfool attacks are slow to compute, and hence 500 adversaries were generated from randomly drawn test images.

Table 2: Classification Accuracy over Adversarial

CIFAR10				
	Fast-FGSM	Halftone	Mixup	SVrandom+
Noise	0.902	0.801	0.907	**0.939**
FGSM	0.866	0.804	0.670	**0.910**
PGD	0.876	0.800	0.456	**0.878**
CW	0.887	0.119	0.904	**0.952**
Deepfool	0.854	0.796	0.925	**0.936**
Average	*0.877*	*0.664*	*0.772*	***0.923***
Change	*0.0%*	*-17.52%*	*-17.34%*	*-3.15%*
CIFAR100				
	Fast-FGSM	Halftone	Mixup	SVrandom+
Noise	0.743	0.492	0.653	**0.750**
FGSM	**0.733**	0.486	0.458	0.683
PGD	**0.736**	0.478	0.362	0.622
CW	0.735	0.027	0.725	**0.771**
Deepfool	0.738	0.506	0.731	**0.767**
Average	***0.737***	*0.398*	*0.586*	*0.719*
Change	*2.36%*	*-19.11%*	*-20.70%*	*-7.82%*
GTSDB				
	Fast-FGSM	Halftone	Mixup	SVrandom+
Noise	0.952	0.928	0.971	**0.973**
FGSM	0.907	0.928	0.855	**0.949**
PGD	**0.928**	0.923	0.719	0.918
CW	0.840	0.012	0.881	**0.935**
Deepfool	0.842	**0.896**	0.887	0.742
Average	*0.894*	*0.737*	*0.863*	***0.903***
Change	*3.35%*	*-21.18%*	*-12.39%*	*-7.86%*

The defense capability of each method is summarized in Table 2. In classification accuracy, our method outperforms the others on all types of attack over CIFAR10 and 3 out

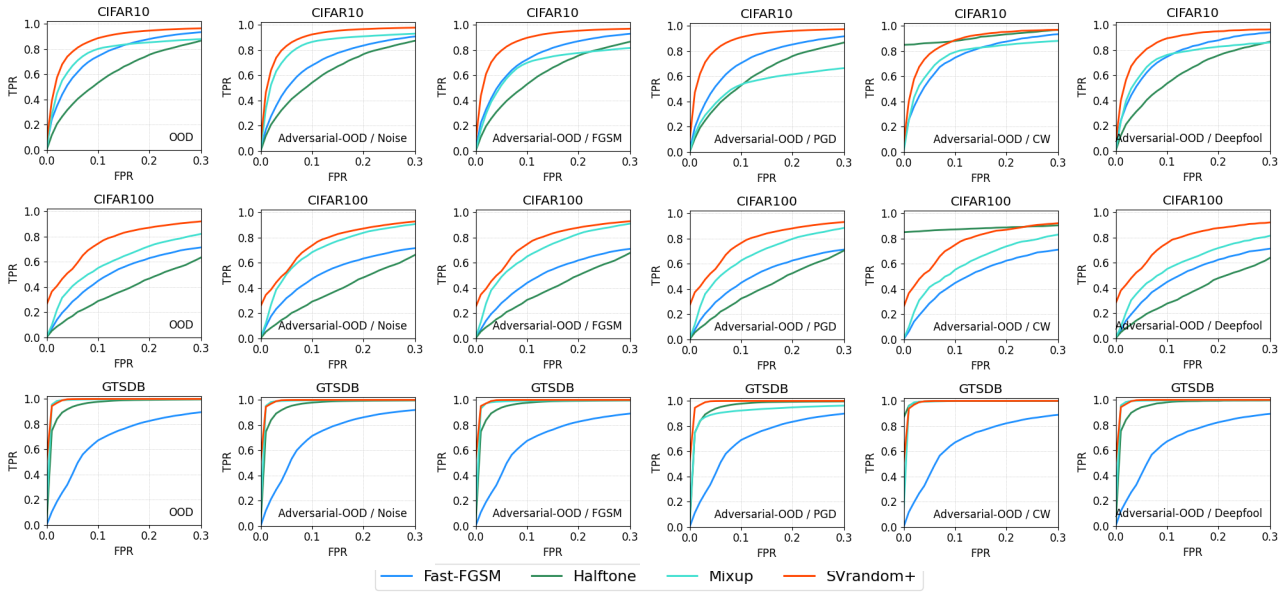

Figure 4: Detection Performance for OOD and Adversarial-OOD Test Sets

of the 5 types of attack on the other two classifiers. On an average, our method achieves the highest accuracy over CIFAR10 and GTSDB, and is very close (0.719) to the Fast-FGSM (0.737) over CIFAR100 .

Since the Fast-FGSM method is specifically trained using FGSM attacks with a large perturbation budget of $8/255$, we can see that the accuracy of CIFAR100 and GTSDB classifiers increase by 2.36% and 3.35% on an average when compared to benign situations (*i.e.*, the accuracy reported in Table 1b). However, our method is still more robust to attacks than both the Halftone and Mixup methods on an average. The Mixup method is also data augmentation based, but it is much more vulnerable to different types of attacks.

Unexpectedly, under Deepfool attack, Halftone and Mixup methods are more robust than the fast-FGDM method over GTSDB. GTSDB includes traffic sign images that are much less complex than the CIFAR datasets. This suggests that the dataset is also an important factor in designing experiments for unbiased evaluation of attack and defense algorithms.

OOD Detection We now apply the blind-test to measure detection capabilities over OOD and adversarial-OOD. OOD scores are computed with Equation 3b and detection thresholds are found using the extended *baseline* as explained in the Proposal section. For SVrandom+, the input parameters *limit* in Algorithm 1 are 10 (reduction) and 18 (swap). The parent classes[1] for CIFAR100, CIFAR10, and GTSDB are 20, 2 and 5, respectively. Details of the parent-child class relationships are given in Appendix (b**??**). The α value in Equation 3b is 0.8 (CIFAR100) and 0.95 (CIFAR10 and GTSDB) for SVrandom+ classifiers, and 1 for all other classifiers that are trained by single-task.

[1]Classification accuracy on parent classes are CIFAR100 86.8%, CIFAR10 99.3%, and GTSDB 99.8%.

The detection performance is summarised in Figure 4. In each subplot, the ID set that a classifier was trained on is shown in the title. The performance of all methods is compared by each type of OOD that is mentioned within the subplot. The horizontal axis is FPR over the ID test set. The vertical axis is a macro average of TPR over all the corresponding OOD test sets listed in Table 1a or their adversaries, i.e., adversarial-OOD. The same tool and parameters described in the previous sub-section were used to generate the adversarial-OOD samples.

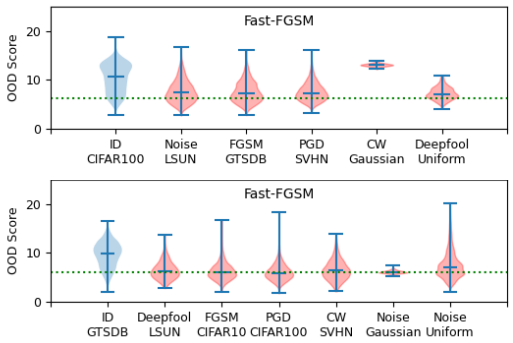

Figure 5: Breakdown of OOD Scores by test sets. The dotted lines are detection threshold at 10% FPR of ID.

Several observations can be made from Figure 4. SVrandom+ outperforms all the other methods in detection performance irrespective of the OOD type, except for the Halftone method on OOD perturbed with the CW attack and the two CIFAR test sets. Most of the images processed with the halftoning operation and subsequently perturbed by the CW attack led to a logit value explosion. This also explains why the classification accuracy of the Halftone method over adversaries degrades the most, as shown in Table 2 (2.7% over CIFAR100 under the CW attack).

The Fast-FGSM method has a mixed impact on OOD detection. Compared to the Mixup method, we can see that more adversarial-OOD samples with FGSM and PGD attacks are detected over CIFAR10. However, it performs the worst over GTSDB and CIFAR100 disregard of OOD type. A breakdown of OOD scores by test sets, shown in Figure 5, reveals that the Fast-FGSM method fails to detect all Gaussian samples perturbed with CW attack as OOD over CIFAR100. Over half of such OOD samples are missed out over GTSDB. This explains why the average TPR of Fast-FGSM is lower than Mixup in most subplots in Figure 4. The reason behind is that the FAST-FGSM method flats out the distributions of both ID and OOD test samples in a way that pushes them both towards each other. In contrast, our method pulls the distributions of ID and OOD apart. As shown in Figure 6.

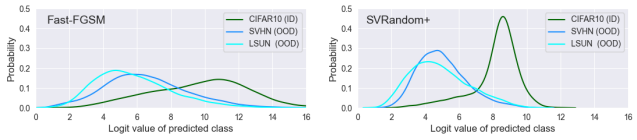

Figure 6: Distribution of logit values of predicted classes by Different Methods

Fail-Safe Control In engineering, fail-safe is a common system design approach for preventing unsafe consequences. Recently, this approach has been explored to mitigate the uncertainty in DNN (Biondi et al. 2019, Weiss and Tonella 2021). We apply OOD detection here as a fail-safe controller to DNN classifiers. Specifically, a classifier only produces class prediction over inputs deemed to be ID.

To assess OOD detection's capability as a fail-safe controller for classifiers in the open world setting, we feed each classifier with a combination of one ID test set, five types of corresponding adversaries, respective OOD test sets from Table 1a and five types of adversaries for each OOD test set. So, a CIFAR10 classifier has 36 test sets, and a GTSDB classifier has 42 test sets in total.

Classification error without fail-safe control is shown in Table 3a. After a fail-safe controller rejects inputs detected as OOD, which otherwise tend to contribute to error predictions, it is not a surprise that the classification error rate reduces across all the methods. However, our method outperforms all others by a significant margin.

Summarising the above three sets of tests, we can see OOD detection is helpful for classifiers to make fewer error predictions irrespective of the defensive training method. We can also observe a persistent conflict between classifiers' robustness to adversaries versus accuracy and the discrimination power in their logits for OOD detection, be it gradient, input transformation, or data augmentation based defensive training techniques. Experiment results indicate our defensive training method overcomes such conflict.

Conclusion

In this paper, we investigated the interplay among adversarial defense, classification accuracy and OOD detection.

Table 3: Overall Performance

Classification Error Rate				
	Fast-FGSM	Halftone	Mixup	SVrandom+
CIFAR10	0.854	0.885	0.867	0.845
CIFAR100	0.878	0.931	0.898	0.878
GTSDB	0.873	0.890	0.874	0.869

(a) Only Adversarial Defense

TPR of OOD Detection @ 10% FPR of ID				
	Fast-FGSM	Halftone	Mixup	SVrandom+
CIFAR10	0.725	0.591	0.742	**0.892**
CIFAR100	0.451	0.393	0.602	**0.750**
GTSDB	0.682	0.982	0.984	**1.000**
Classification Error Rate				
	Fast-FGSM	Halftone	Mixup	SVrandom+
CIFAR10	0.238	0.355	0.229	**0.093**
CIFAR100	0.478	0.542	0.360	**0.227**
GTSDB	0.280	0.016	0.023	**0.002**

(b) Adversarial Defense and OOD Detection

We proposed a singular subspaces based data augmentation technique and a multi-level semantics based OOD detection metric to reduce the classifiers' sensitivity to adversarial attacks. Our method overcomes the performance trade-off between OOD detection and classification robustness. We conducted an extensive evaluation that aligns with situations in the open world. Comparison experiments showed that our method is the most robust among the four defensive training methods.

References

Ahmed, F.; and Courville, A. 2020. Detecting semantic anomalies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 3154–3162.

Andrews, H.; and Patterson, C. 1976. Singular value decomposition (SVD) image coding. *IEEE transactions on Communications*, 24(4): 425–432.

Biondi, A.; Nesti, F.; Cicero, G.; Casini, D.; and Buttazzo, G. 2019. A safe, secure, and predictable software architecture for deep learning in safety-critical systems. *IEEE Embedded Systems Letters*, 12(3): 78–82.

Bulusu, S.; Kailkhura, B.; Li, B.; Varshney, P. K.; and Song, D. 2020. Anomalous example detection in deep learning: A survey. *IEEE Access*, 8: 132330–132347.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, 39–57. IEEE.

Chang, C.-C.; Tsai, P.; and Lin, C.-C. 2005. SVD-based digital image watermarking scheme. *Pattern Recognition Letters*, 26(10): 1577–1586.

Dong, Y.; Fu, Q.-A.; Yang, X.; Pang, T.; Su, H.; Xiao, Z.; and Zhu, J. 2020. Benchmarking adversarial robustness on image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 321–331.

Gong, C.; Ren, T.; Ye, M.; and Liu, Q. 2021. MaxUp: Lightweight Adversarial Training With Data Augmentation Improves Neural Network Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2474–2483.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Guo, Q.; Zhang, C.; Zhang, Y.; and Liu, H. 2015. An efficient SVD-based method for image denoising. *IEEE transactions on Circuits and Systems for Video Technology*, 26(5): 868–880.

Hendrycks, D.; and Gimpel, K. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Houben, S.; Stallkamp, J.; Salmen, J.; Schlipsing, M.; and Igel, C. 2013. Detection of Traffic Signs in Real-World Images: The German Traffic Sign Detection Benchmark. In *International Joint Conference on Neural Networks*, 1288.

Hsu, Y.-C.; Shen, Y.; Jin, H.; and Kira, Z. 2020. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10951–10960.

Huang, Y.-T.; Liao, W.-H.; and Huang, C.-W. 2021. Defense Mechanism Against Adversarial Attacks Using Density-based Representation of Images. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 3499–3504. IEEE.

Kim, H. 2020. Torchattacks: A pytorch repository for adversarial attacks. *arXiv preprint arXiv:2010.01950*.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31.

Liang, S.; Li, Y.; and Srikant, R. 2018. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *International Conference on Learning Representations*.

Liu. 2017. pytorch-cifar. https://github.com/kuangliu/pytorch-cifar.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; Fawzi, O.; and Frossard, P. 2017. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1765–1773.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; and Frossard, P. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2574–2582.

Shafaei, A.; Schmidt, M.; and Little, J. J. 2018. A less biased evaluation of out-of-distribution sample detectors. *arXiv preprint arXiv:1809.04729*.

Shafahi, A.; Najibi, M.; Xu, Z.; Dickerson, J.; Davis, L. S.; and Goldstein, T. 2020. Universal adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 5636–5643.

Shorten, C.; and Khoshgoftaar, T. M. 2019. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1): 1–48.

Song, L.; Sehwag, V.; Bhagoji, A. N.; and Mittal, P. 2020. A critical evaluation of open-world machine learning. *arXiv preprint arXiv:2007.04391*.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Weiss, M.; and Tonella, P. 2021. Fail-safe execution of deep learning based systems through uncertainty monitoring. In *2021 14th IEEE Conference on Software Testing, Verification and Validation (ICST)*, 24–35. IEEE.

Wong, E.; Rice, L.; and Kolter, J. Z. 2019. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*.

Zenati, H.; Foo, C. S.; Lecouat, B.; Manek, G.; and Chandrasekhar, V. R. 2018. Efficient gan-based anomaly detection. *arXiv preprint arXiv:1802.06222*.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; and Jordan, M. 2019. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 7472–7482. PMLR.

Zhang, Y.; and Yang, Q. 2021. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*.

Appendix: Parent-Child Class Relationship

In CIFAR100, its original 20 super-classes were used as parent classes. Four classes of man-made objects in CIFAR10 are merged into one parent group and the remaining six animal classes into another parent group. By shape, color, and content, the 43 traffic sign classes in GTSDB are grouped into 5 parent classes, as shown in the 5 black-bordered boxes in Figure 7.

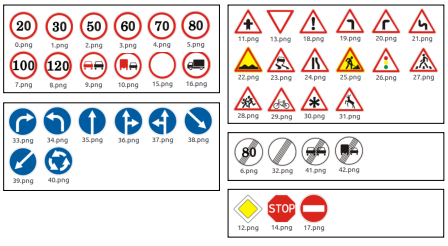

Figure 7: Parent classes of GTSDB in Meta Images